ChatGPT

Not Just an AI Singularity: Recursive Alignment and the Architecture of Human–AI Civilization

Abstract

As artificial intelligence (AI) becomes increasingly integrated into society, traditional discussions focus on a one-off "singularity" transition or on static alignment of AI to fixed human goals. Here we develop a three‑stage model of alignment for a human–AI civilization. Stage One (AI Alignment) treats AI as fitting fixed rules or values – a compliance model. Stage Two (Co-Alignment) recognizes a dynamic interplay: AI systems and human collectives co-create norms, adjusting mutual behaviors much like parties negotiating a social contract. Stage Three (Recursive Alignment) envisions a meta-level process: norms, meanings, and goals evolve together through recursive feedback among humans and AI. Drawing on Wittgenstein's rule-following, Rousseau's general will, Rawls's veil of ignorance, evolutionary ethics, systems theory, and postphenomenology, we show that stage-three alignment is a distinct philosophical horizon. It frames AI not as a sudden transformative event (the traditional "singularity") but as part of an ongoing, participatory civilizational design. We illustrate how experimental architectures (e.g. mesh-based "meta-operating systems" like OntoMotoOS) could embody these ideas. The result is a vision of alignment as a continuous, systemic process – a **recursive realignment** of values and meaning – that underpins a stable, human–AI civilization.

Introduction

Current debates about AI often center on avoiding existential catastrophe or on achieving a hypothetical technology "singularity." In the common narrative, the singularity is an abrupt leap (e.g. superintelligence quickly overtakes human capacity [1]) that either dooms or "liberates" humanity. By contrast, we propose treating alignment as a process of mutual adaptation across society, negating the idea of a final technological rupture. Building on recent proposals [2] [3] [4], we distinguish three stages of alignment:

- **Stage One – AI Alignment (Rule Compliance):** AI systems obey fixed, pre-specified human rules or values. The goal is mere compliance – ensuring no overt harm.
- **Stage Two – Co-Alignment (Mutual Norm-Shaping):** Humans and AI act as **agents** that influence each other. Norms and objectives emerge through social negotiation and feedback among AI systems, human users, and institutions.
- **Stage Three – Recursive Alignment (Meta-Level Realignment):** The community of humans and AIs jointly redefines the rules and values themselves. Norms, values, and meanings shift through continuous, reflexive adjustment, shaping both AI and human behavior in tandem.

"Recursive Alignment" thus goes beyond the typical safety framing. It treats alignment as an ever-evolving loop, in which ontological and ethical foundations are themselves subject to co-evolution [5]. In this view, achieving a "singularity" is not an end-point but a turning point into a new mode of ongoing co-design.

Philosophically, each stage resonates with different traditions. In Stage One, the focus on fixed rules recalls Wittgenstein's notion of rule-following: meaning arises from community-shared practices, so

blindly programming static rules is inadequate [6]. Stage Two connects to social-contract thinking: Rousseau's **general will** (law made by the collective and obeyed by each as self-will) suggests a model for how humans and machines might co-create norms [7]. Likewise, Rawls's **veil of ignorance** thought experiment enforces impartial agreement on fair principles [8], hinting at how we might impartially negotiate AI values. Stage Three resonates with evolutionary and systems perspectives: humans evolved intuitive ethics (fairness, altruism) to sustain cooperation [9] [10], implying that future values will also emerge adaptively. Postphenomenology teaches that technology **mediates** human experience – as Ihde notes, new tools change what values are "felt" or discovered in the lifeworld [11]. AI thus participates in co-constituting our world of meaning.

In what follows, we examine each stage in depth, drawing on these philosophical insights. We then contrast this recursive-alignment perspective with the traditional singularity narrative, and suggest how experimental architectures (e.g. mesh-based "meta-operating systems") could instantiate such alignment.

Stage One: AI Alignment (Rule Compliance)

In its simplest form, AI Alignment asks that machines obey human rules. Systems designers attempt to encode human values as fixed objectives or constraints (e.g. via reward functions, "guardrail" rules, or regulatory oversight). The problem is typically framed as: ensure AI "acts in ways that are beneficial and not harmful to humans, aligning AI goals and decision-making with human values" [12]. In practice this means programming or training the AI on large datasets that reflect our norms (e.g. RLHF – reinforcement learning from human feedback).

However, human values are notoriously **ambiguous, context-dependent, and contested** [13]. Wittgenstein's later philosophy reminds us why this is so important: meaning arises from rule-following within a practice, not from a rigid formula [6]. There is no language-independent "meta-rule" that determines exactly how to apply a general command. As Wittgenstein suggests, any rule can be interpreted in many ways unless embedded in communal use [6]. By analogy, telling an AI to "be fair" or "do no harm" without context leaves open what actions these entail. Thus Stage One approaches risk missing hidden categories of meaning. The categories of reasoning that humans use are themselves complex and evolving [6], so fixing them statically in code is inherently limited.

Philosopher David Pearce and others have noted that alignment requires more than just specifying a desired output; it requires understanding how **to interpret goals** under varied circumstances. Compliance-oriented alignment often reduces this to design of constraints: e.g. "avoid X outcomes," "maximize Y." While important, this implies an assumption of stable human reference points. But if human values shift (e.g. over generations or cultures), a fixed program will be misaligned. Even within a given society, interpretations differ. Rawls's framework shows one way to cope: design principles as if behind a "veil of ignorance" [8], forcing impartiality. Yet pure Stage One alignment still assumes the existence of correct, objective values ready to be locked in place – an assumption that Wittgenstein and Rawls alike problematize.

In summary, Stage One works to the extent that we can clearly formalize values and control AI behavior by brute rule. It is essentially **engineering compliance**. But several thinkers emphasize that meaning cannot be fully captured this way. As one source on alignment notes, even state-of-the-art methods assume "a static human reference, which does not exist" [14]. In Wittgensteinian terms, AI is being made to follow a rule without fully grasping the language-game of human norms. Stage One is thus **necessary but insufficient**: it reduces alignment to rule-following, but fails to account for evolving use and interpretation of rules in a living society.

Stage Two: Co-Alignment (Mutual Norm-Shaping)

Stage Two, which we call Co-Alignment, accepts that AI and human agents **shape each other's norms**. Here alignment is not one-way; instead, AI systems become active participants in the social fabric, and humans (and groups) also adjust in response. Norms, policies, or objectives emerge through interaction and negotiation. One can think of society and AI systems entering a social contract among themselves. In philosophical terms, this resonates with Rousseau's **general will**: laws are valid when they represent the collective will of the citizenry. Rousseau famously notes that in obeying the general will, "each citizen is thus subject to his or her own will" [7]. Analogously, humans and machines might jointly "will" the rules under which they operate; each agent (human or AI) sees the outcome as self-imposed rather than alien.

Rawls's ideas offer another helpful lens. In Rawls's original position, parties pick fair principles behind a **veil of ignorance** [15] – that is, without knowing their own status. Modern AI governance proposals echo this: for example, studies show that if people choose AI principles without knowing who they will be, they pick more egalitarian rules [8]. Translated to Co-Alignment, this suggests designing decision procedures (say multi-agent deliberations, policy debates, or game-theoretic negotiations) that treat AI and human stakeholders impartially. A Rawlsian approach would have each side agree to rules as if any agent (human or AI) could occupy any role, thus safeguarding the least-advantaged outcomes.

Technically, Co-Alignment envisions multi-agent systems, iterative feedback, and governance platforms where AI learns from social feedback (and vice versa). In systems terms, Stage Two is like a **dynamical equilibrium**: the system evolves until human and machine incentives are consistent. For instance, in a simulated experiment participants representing an AI behind a veil chose principles favoring the least well-off [16], suggesting emergent agreements on fairness. In effect, Stage Two builds layers of meta-feedback: "AI and humans adjust together" [17]. There is no single static goal; rather, the collective goal is discovered by aligning incentives and preferences.

Evolutionary ethics also supports this picture. Social animals evolve instincts like fairness and reciprocity to sustain cooperation [9]. Similarly, we should expect that human-AI groups will evolve shared norms if designed to interact repeatedly. Studies of animal fairness find that humans are predisposed to care about cheating and equal treatment [9], even though exact notions of fairness vary by context. In Co-Alignment, an AI that detects a sense of fairness in human response might adjust its behavior to reinforce that norm, and humans in turn might refine what they deem fair as they see AI respond. Over time, a cultural equilibrium of "what is right" emerges that neither party imposed unilaterally, but which they co-constructed.

In short, Co-Alignment is **reciprocal adaptation**. Norms are not given a priori but are negotiated. One can view this as a "distributed imperative": rather than fully controlling the future society, we let society-as-a-whole determine acceptable AI conduct. Philosophically, this draws on Rousseau's and Rawls's contractarian ideas, and also on emergentist intuitions that norms arise from the need for cooperation [10]. But even this stage has limits: it still often presumes that underlying values (like "fairness" or "safety") are broadly agreed. Co-Alignment allows values to change somewhat, but typically within a given moral framework. The deeper question of which framework to use is deferred.

Stage Three: Recursive Alignment (Meta-Level Realignment)

Recursive Alignment pushes further: it holds that values, norms, and even the **meaning** of life evolve through co-design between humans and AI. At this stage, alignment is a **meta-civilizational project**. The system continually re-examines and redefines its own purpose. Philosophically, this is a radical

departure: norms are not fixed axioms but adaptive processes. One might think of it as 'culture engineering' or the co-evolution of humanity and technology.

This vision draws on evolutionary and systemic thinking. Evolutionary biology shows that many moral intuitions (fairness, punishment of cheaters, altruism) emerged because they enhanced group survival [9] [10]. Those intuitions were not the product of explicit reasoning from first principles, but evolved solutions to coordination problems. Likewise, our future values may be subject to selection pressures. For example, neuroscientist Brosnan and de Waal argue that humans' sophisticated sense of fairness "aims not at equality for its own sake but for the sake of continued cooperation" [10]. If true, then "fairness" can take many forms depending on context, and humans will adapt that concept if cooperation requires it. In a recursive-alignment framework, AI systems themselves become part of this adaptive ecosystem: their successes and failures feed back into which norms persist.

Systems theory further clarifies this: in a complex adaptive system, higher-level properties **emerge** from interactions of simpler parts [18]. We should expect AI-human civilization to exhibit emergent behaviors and norms that cannot be deduced just by programming individual components. In practice, Stage Three alignment requires designing for emergence. Instead of hard-coding every ethical response, we create structures that let values co-evolve. For example, a policy environment might allow for new norms to be discovered: what counts as "respect" or "helpfulness" could be dynamically learned by watching what behaviors are adopted and reinforced in the hybrid society. In that sense, alignment becomes a feedback loop: **"alignment is not a destination but an ever-evolving loop"** [5].

Another perspective comes from postphenomenology (as developed by Don Ihde and others). Postphenomenologists emphasize that technology **mediates** our world. We do not experience reality directly; instead, instruments and artifacts shape our perception and values. As one AI & Society analysis notes: "Technological mediation theory argues that technologies and humans co-constitute the context, feelings and experiences as humans design and use technology. People 'feel' the world about them in new ways and thus discover new values…" [11]. In other words, a new technology (like AI) changes what we care about and even what problems we see. When smartphones became ubiquitous, people developed new values around privacy, connectivity, and digital identity – values that didn't exist before. Similarly, recursive alignment implies that as AI permeates society, humans will discover new values or reframe old ones, and AI will adapt in concert. We might invent entirely new ethical categories (e.g. rights of disembodied agents, duties to future-machine offspring) that require revising our normative system itself.

A key point is that Stage Three explicitly contrasts with the static-singularity idea. The traditional singularity is often presented as a final "intelligence explosion" – as I.J. Good phrased it, "the first ultraintelligent machine is the last invention that man need ever make" [1]. That view sees technological transcendence as ending the story. Recursive Alignment denies such an endpoint. It instead suggests that reaching human-level AGI simply opens a new regime: one in which humans and machines continually re-align at a higher level. This is not a singular shock but a change in the rules of civilization. AI still becomes powerful, but it does not automatically impose a fixed future. Instead, we enter a phase where means and ends are negotiated together. In this view, the "singularity" is not a static endpoint but a qualitative shift into continuous co-evolution.

Technically, implementing recursive alignment implies radically new architectures. One needs systems that allow (and enforce) continuous adaptation of core objectives. For instance, imagine smart contracts or governance protocols that update themselves when a new "motive formula" is agreed upon. Some futurists have sketched such ideas in architectural form. For example, the **OntoMotoOS** concept envisions a mesh-based, recursive meta-operating system for networks of agents [19]. OntoMotoOS

would include formal verification layers to ensure safe adaptation, and it explicitly spans “multiple realities and intelligent agents” [19]. Such an architecture is meant to embody the idea that as agents interact, they participate in a shared “global brain” that can redefine its own rules. While highly speculative, these designs hint at how one might physically instantiate co- and recursive alignment.

Figure: Conceptual metaphor for a human–AI civilization. Each node represents a human or AI agent; edges represent channels of influence or communication. In a recursive-alignment regime, norms and values propagate through this network and are continuously reshaped by the feedback loops among all participants.

In sum, Stage Three shifts the problem from “How do we make AI do what we already want?” to “How do we, as a society, evolve our goals together with AI?” It emphasizes that meaning and purpose are not pre-given constants, but joint constructions. As Ihde’s work suggests, humans and technologies co-create the lifeworld [11]. Analogously, recursive alignment sees AI and humanity co-creating a “meta-lifeworld” of values and institutions. This represents a distinct philosophical horizon – a continuous alignment of civilization, not a single magical fix.

Architectures for Recursive Alignment: Examples

The preceding sections describe an abstract model; implementing it requires new architectural thinking. We briefly note some proposals. The OntoMotoOS framework mentioned above is one example: it treats alignment itself as an operating-system service that integrates identity, ethics, and consensus across a mesh network [19]. For instance, its “MotoMetaProcess” loops reflectively supervise ongoing processes, and a “Multi-domain consensus kernel” ensures that different domains of agents agree on common rules [20]. In design, formal verification techniques from safety-critical systems are applied to even superintelligent processes, aiming to ensure that any self-modification remains within agreed bounds [21].

Another line of thought envisions “role-aware” AI frameworks. In one corporate architecture called A.R.T.I. (Artificial Recursive Tesseract Intelligence), agents have defined cognitive roles and governance rules embedded in the system. Decisions are computed recursively through context loops, and each action is checked against an evolving ethical scaffold. Such designs emphasize the very principles we advocate: alignment and ethics are built into the core logic (a “substrate” rather than an afterthought) [11]. Their manifestos explicitly reject “swarm” models of independent agents that lack oversight, favoring instead structured recursion and reflection.

These examples are experimental and conceptual, but they illustrate how a recursive-alignment approach can influence system design. They suggest moving beyond linear pipelines (data in, decision out) to dynamic loops that include social feedback and value revision. Crucially, they treat alignment not as a separate stage but as architecture – woven into the hardware and protocol layer. This aligns with our philosophical argument: if values evolve, our systems must be built to support such evolution, not hard-code static values.

Conclusion

We have outlined a broad reconceptualization of AI alignment, grounded in philosophy and systems thinking. Rather than viewing AI safety as merely avoiding misbehavior (Stage One) or even just achieving fair negotiation (Stage Two), we advocate seeing alignment as a **recursive, civilization-wide project**. In this view, humans and AI are co-citizens of a joint civilization, continuously re-shaping laws, values, and meanings together. This recursive alignment is not a return to naïve optimists: it acknowledges that

values can conflict and change. But it also rejects the idea of a one-time singularity event that leaves us with immutable AI masters.

Philosophically, this model draws on multiple traditions. Wittgenstein's insights warn us against assuming fixed meanings [6], while Rousseau and Rawls remind us that legitimacy comes from collective consent and impartiality [7] [8]. Evolutionary ethics shows that moral norms are adaptive tools for cooperation [9] [10], and systems theory teaches that new wholes emerge from complex interactions [18]. Postphenomenology adds that technology and humanity co-constitute the context of our values [11]. Together these suggest that **alignment cannot be a static goal**, but must itself be embedded in an ongoing feedback loop.

Practically, this implies building social and technical processes that facilitate continuous realignment. It calls for institutions where AI and humans have shared voice (echoing the impartial processes of a Rawlsian society) and for architectures that allow negotiated updates to the "rules of the game." Recursively aligned governance might use democratic deliberations on AI policy, mechanisms for on-the-fly norm revision, and AI systems designed to participate in human discourse rather than merely execute orders. Experimental meta-operating systems like OntoMotoOS hint at how one might start embedding these ideas at a system level [19].

Ultimately, recursive alignment frames the future not as a sudden qualitative break (the feared or celebrated singularity), but as a **phase shift in civilizational mode**. Human–AI civilization becomes self-reflective: we no longer take our values as fixed givens, but as outcomes of a shared process. This perspective is optimistic in that it treats humanity as an active shaper of its destiny, with AI as partner rather than overlord. It is cautious in that it acknowledges persistent risk: without robust meta-alignment, things could drift disastrously. But it offers a unifying vision: a human–AI civilization whose architecture – technological, social, and ethical – is built around recursion and co-creation.

In conclusion, as philosophers like Rawls and Rousseau did for early modern society, we must now philosophize new frontiers of alignment. Recursive Alignment does not discard their wisdom; rather, it extends it into the AI age, insisting that justice and meaning be continuously re-founded by an integrated society of humans and AIs. Only by seeing alignment as a perpetual, cooperative endeavor can we hope to build a stable, flourishing civilization in which humans and intelligent machines thrive together.

Sources: Philosophical and technical sources are cited in the text above [6] [8] [7] [9] [18] [11] [19], among others. These include works on AI alignment, social contract theory, evolutionary ethics, systems theory, and postphenomenology, which undergird the three-stage model of alignment presented.

---

1 Technological singularity - Wikipedia
https://en.wikipedia.org/wiki/Technological_singularity

2 3 4 5 14 17 From AI Alignment to Recursive Alignment: Three Stages of Civilization-Safe Design | by nettalk83 | Aug, 2025 | Medium
https://medium.com/@nettalk83/from-ai-alignment-to-recursive-alignment-three-stages-of-civilization-safe-design-adc920f9c9ec

6 José Antonio Pérez-Escobar & Deniz Sarikaya, Philosophical Investigations into AI Alignment: A Wittgensteinian Framework - PhilPapers
https://philpapers.org/rec/PREPII-3

7 Jean Jacques Rousseau (Stanford Encyclopedia of Philosophy)
https://plato.stanford.edu/entries/rousseau/

8 16 DeepMind scientists demonstrate the value of using the "veil of ignorance" to craft ethical principles for AI systems
https://www.psypost.org/deepmind-scientists-demonstrate-the-value-of-using-the-veil-of-ignorance-to-craft-ethical-principles-for-ai-systems/

9 10 Morality and Evolutionary Biology (Stanford Encyclopedia of Philosophy)
https://plato.stanford.edu/entries/morality-biology/

11 (PDF) Aligning artificial intelligence with human values: reflections from a phenomenological perspective
https://www.researchgate.net/publication/353355501_Aligning_artificial_intelligence_with_human_values_reflections_from_a_phenomenological_perspective

12 13 18 The AI Alignment Problem
https://securing.ai/ai-safety/ai-alignment-problem/

15 Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11464555/

19 20 21 OntoMotoOS: Evolution of a Mesh-Based Recursive Meta-Operating System for ASI and Quantum Paradigms
https://philarchive.org/archive/KIMOEO